HUMMING TRANSCRIPTION

FINAL PROJECT FOR IFT-7030

Machine Learning for Signal Processing

Isaac Neri Gomez Sarmiento

Faez Amjed Mezdari

Shubham Gupta

Outline

- Introduction
- Challenges
- Metrics overview
- Classical methods
- HMM based methods
- Midi to musical notes representation
- CNN based methods
- Conclusion

What is humming transcription?

Humming (Fredonner en français)

Recorded sound signal

Music representation

AMT (Automatic Music Transcription) has largely focused on instrumental data

Not much research on AMT for vocals

Collection of humming data is hard

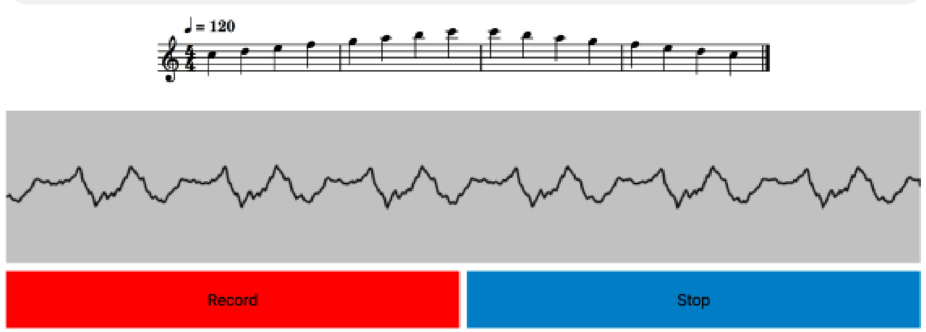
Humming data is noisy – humans dont have same direct control over producing specific notes.

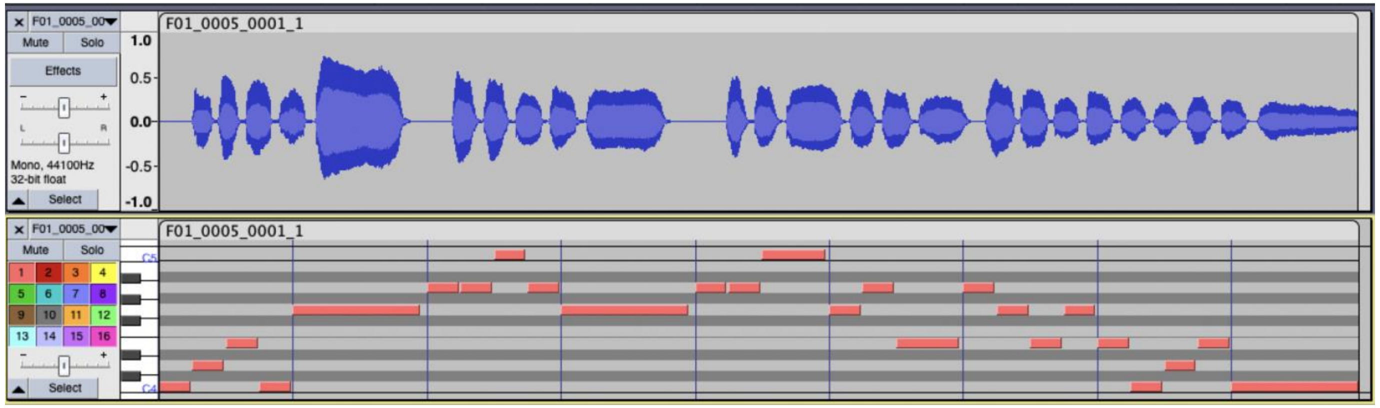
Applications:

- Hum to search on google/spotify
- Write a composition just by humming

HumTrans Dataset

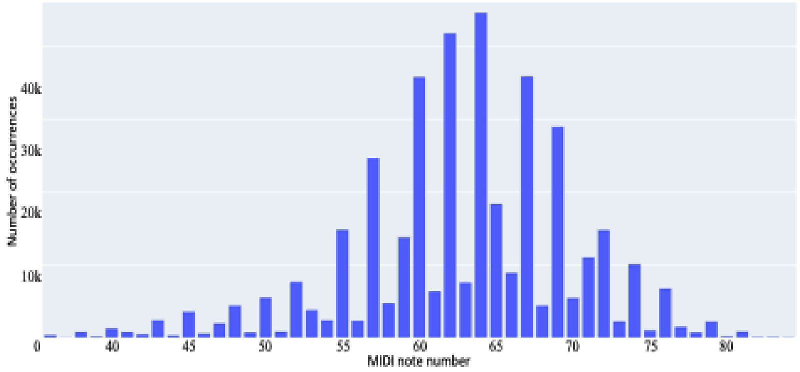
- Largest dataset consisting solely of hummed melodies, released in Sep 2023
  - 1000 music segments
  - 10 college students proficient with music
  - 44KHz
  - 56.22 hours
  - 14,614 total recordings

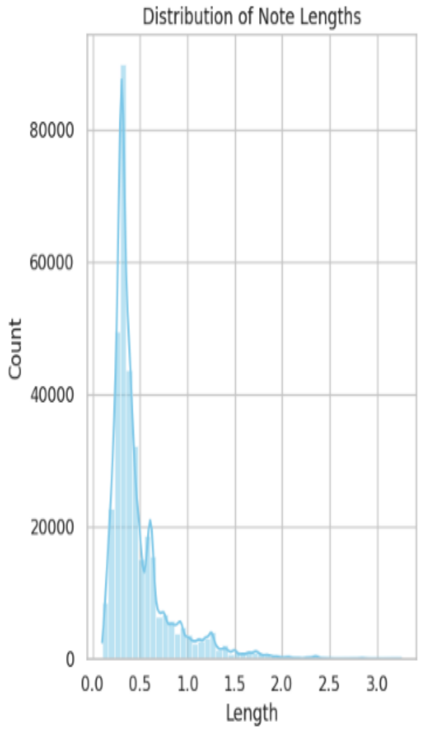


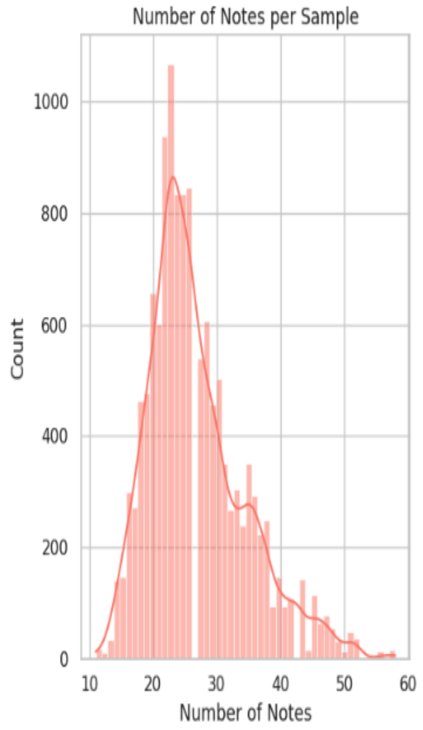


HumTrans Dataset

- 60 unique pitches
- 27 notes per sample on avg
- 500ms avg note length
- 26 – lowest midi pitch
- 88 – highest midi pitch







MIDI number

$$p = 69 + 12\log_2\left(\frac{f}{440}\right) \qquad \begin{array}{c} p \in [0,127] \\ f \in [8.2, 12543.9]\ Hz \end{array}$$

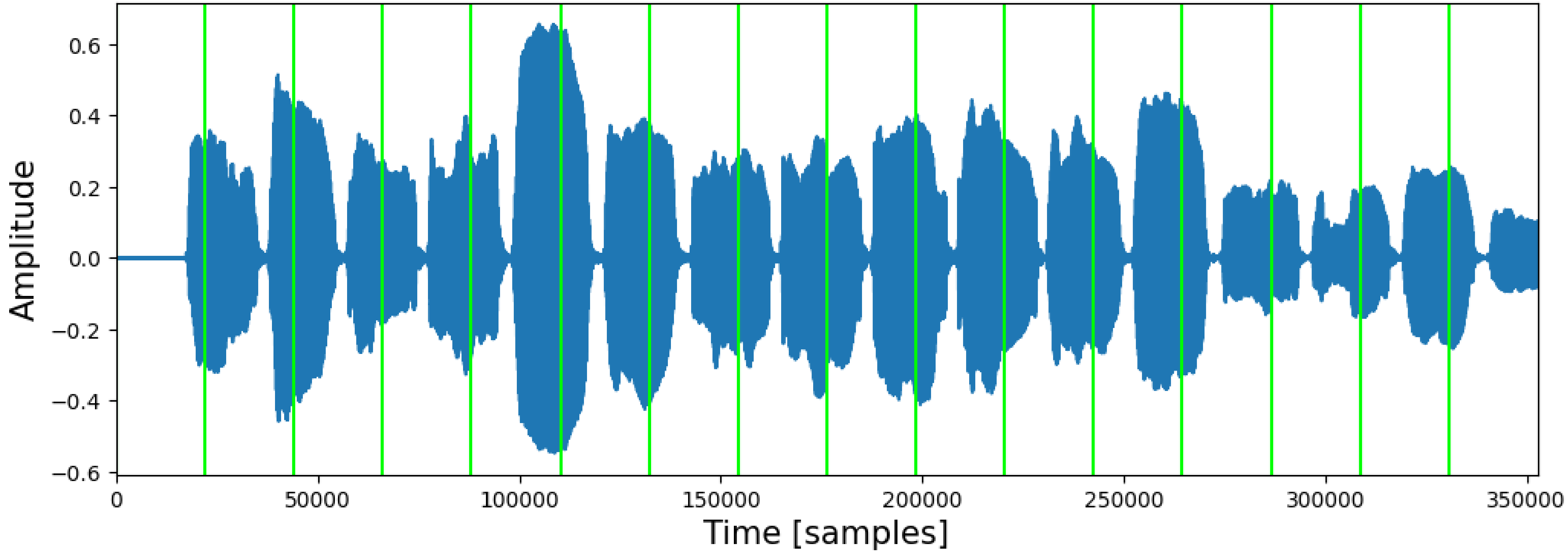
Challenge of the dataset

- Ground truth onsets and offsets are not well aligned

(a.k.a pseudo-ground truth)

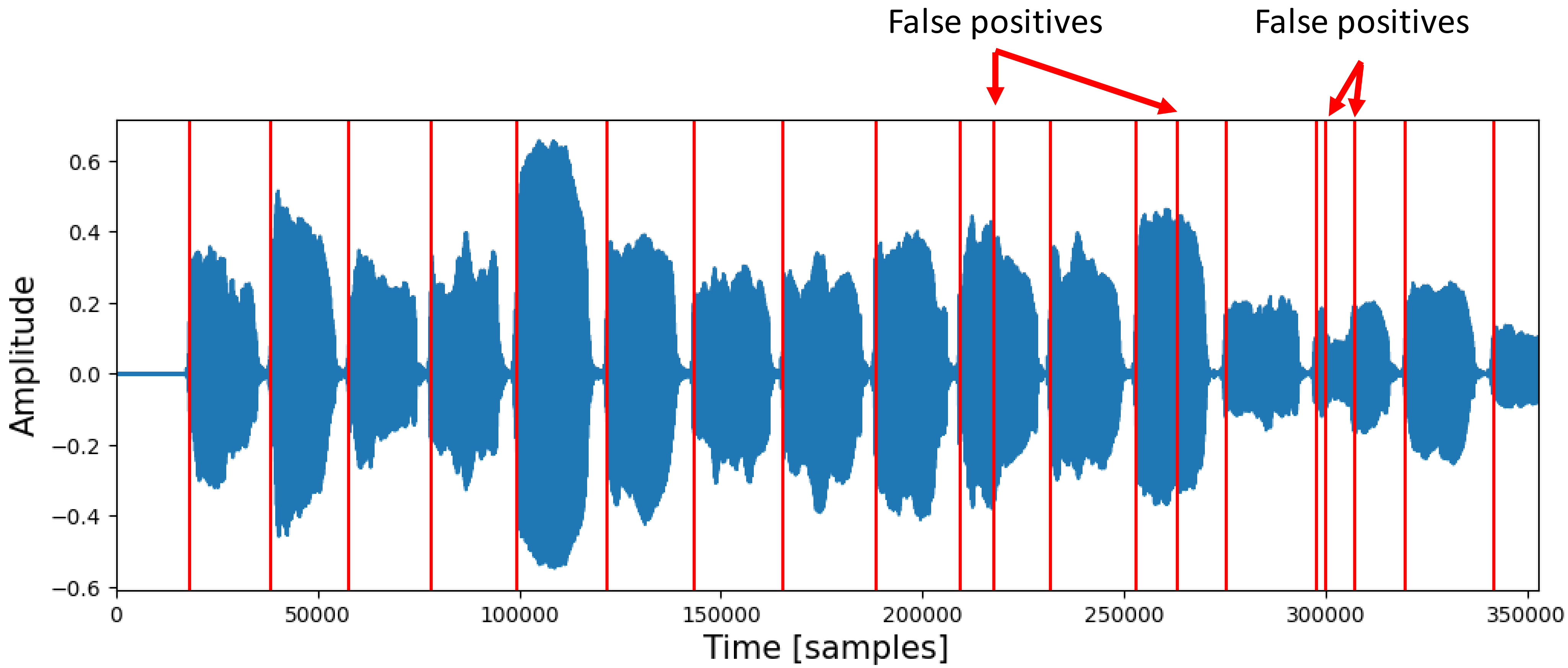
Wav sound

Midi sound



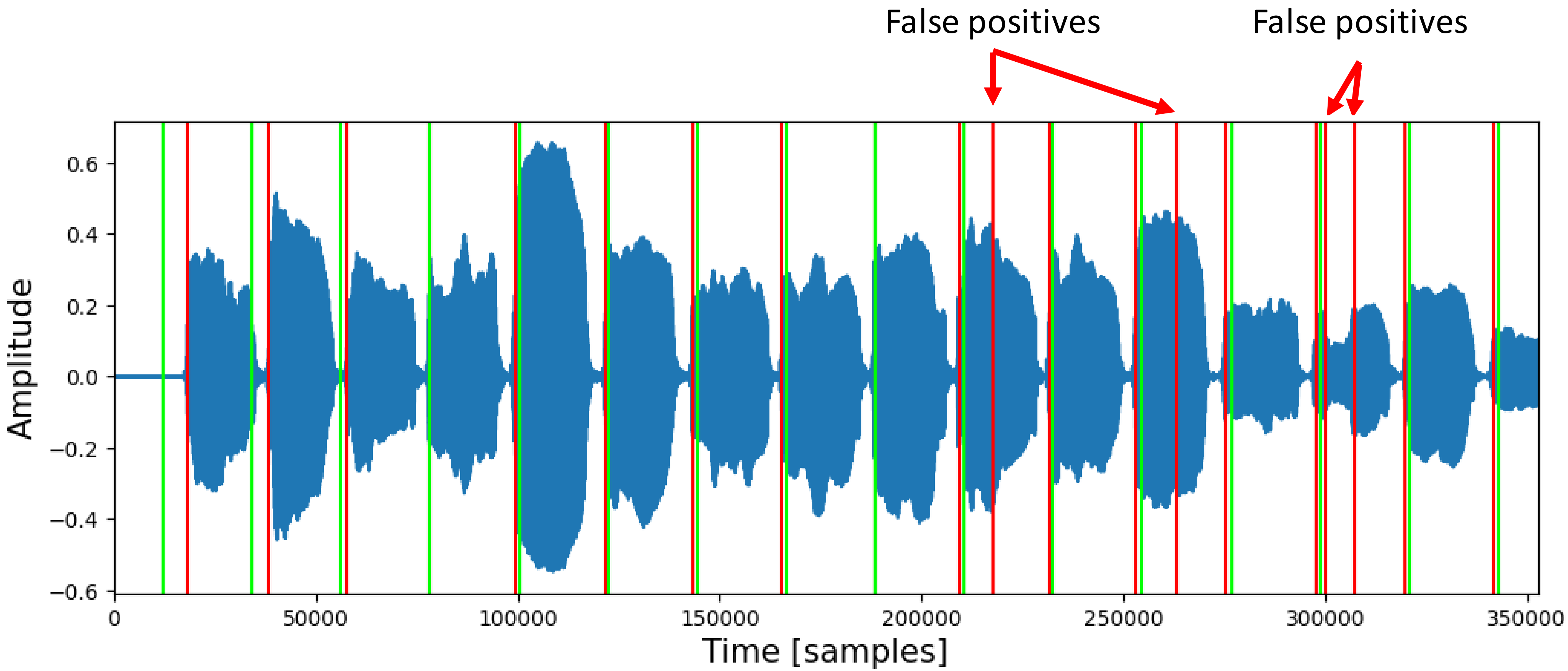
Challenge of the dataset

- Let’s try Librosa’s onset detection method
- It looks for peaks when there is a change in frequency



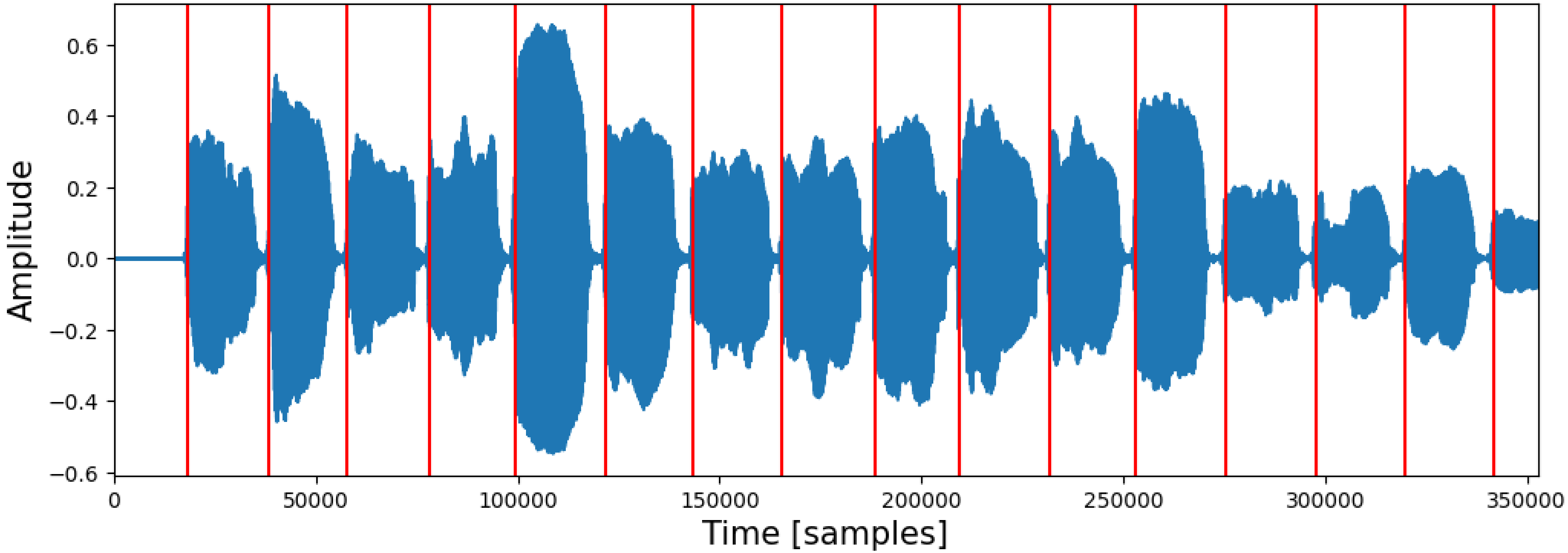
Challenge of the dataset

- Correcting onsets: Align pseudo ground truth to Librosa's onsets, by aligning their centroids and then match closest onsets.



Challenge of the dataset

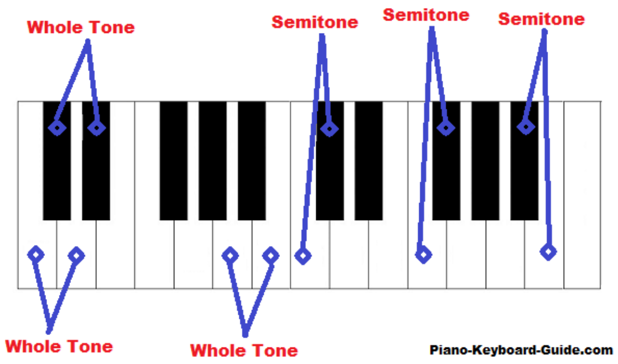
- Now we have better onset estimations ☺!



Some music theory

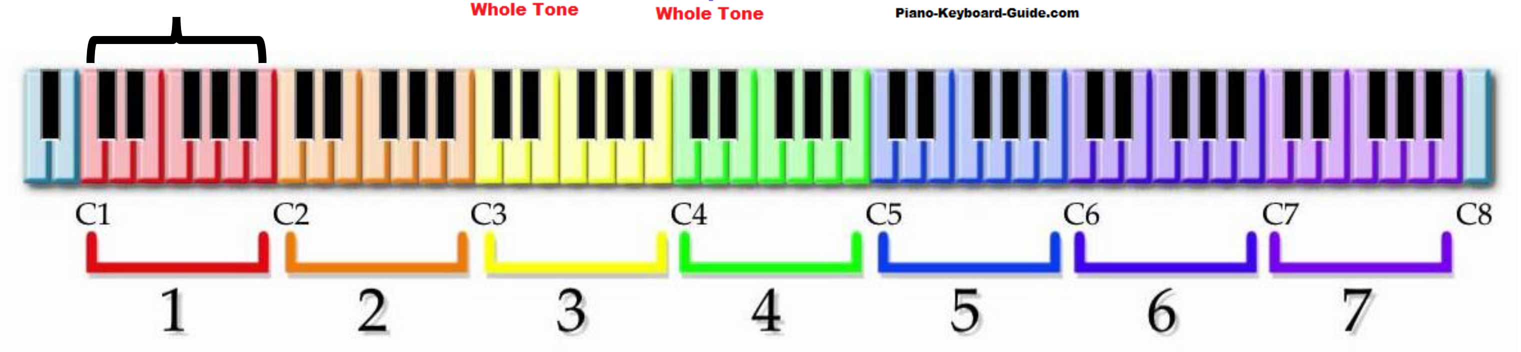
- Piano has 12 semitones for each octave

You can have the same melody at different octaves (low-high pitched)



Original midi

12 keys



1 octave lower midi

Octaves

1 octave higher midi

Piano has 88 keys → 7 octaves +3 lower notes.
8 white keys from [C_i, C_{i+1}]

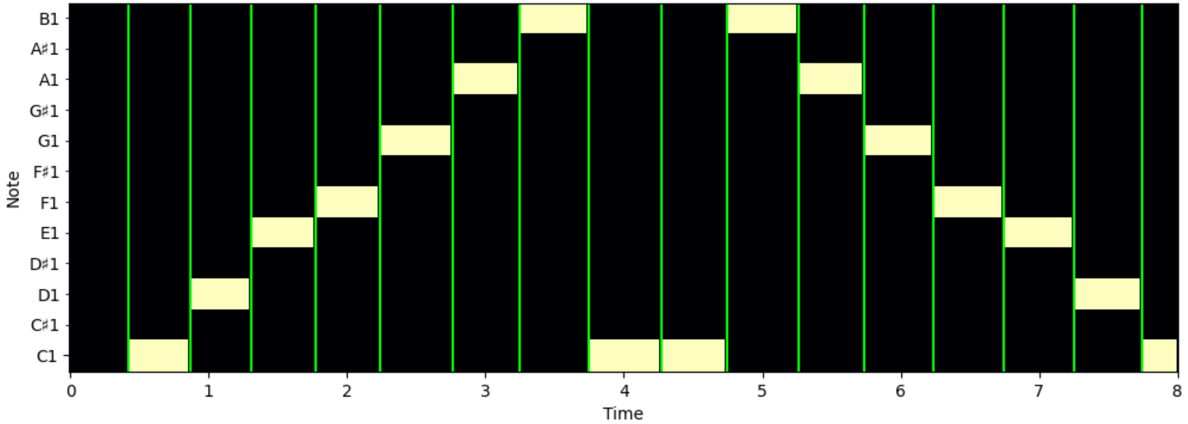
Metrics

- Percentage of correct notes in a file – mean, std-dev
- Percentage of notes in test set predicted correctly
- Percentage of files in test set predicted correctly

- Octave Invariant
- Octave Aware

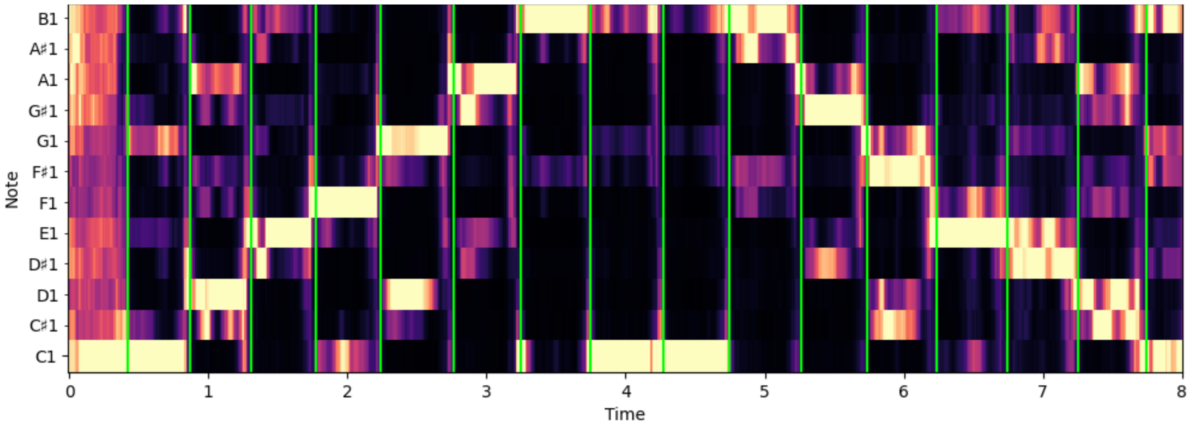
Pitch estimation with chroma-based features

- Onset Estimation using Librosa.
- Pitch Estimation given onsets:
  - Project frequency spectrum onto 12 bins (12 semitone pitch classes), regardless of octave.



Midi ground truth

Octave invariant ground truth
mod(MIDI, 12)



Estimated midi (prediction)

Pitch estimation with chroma-based features

TEST SET

Accuracy=64%

Precision=57%

Recall= 62 %

F1-Score=58%

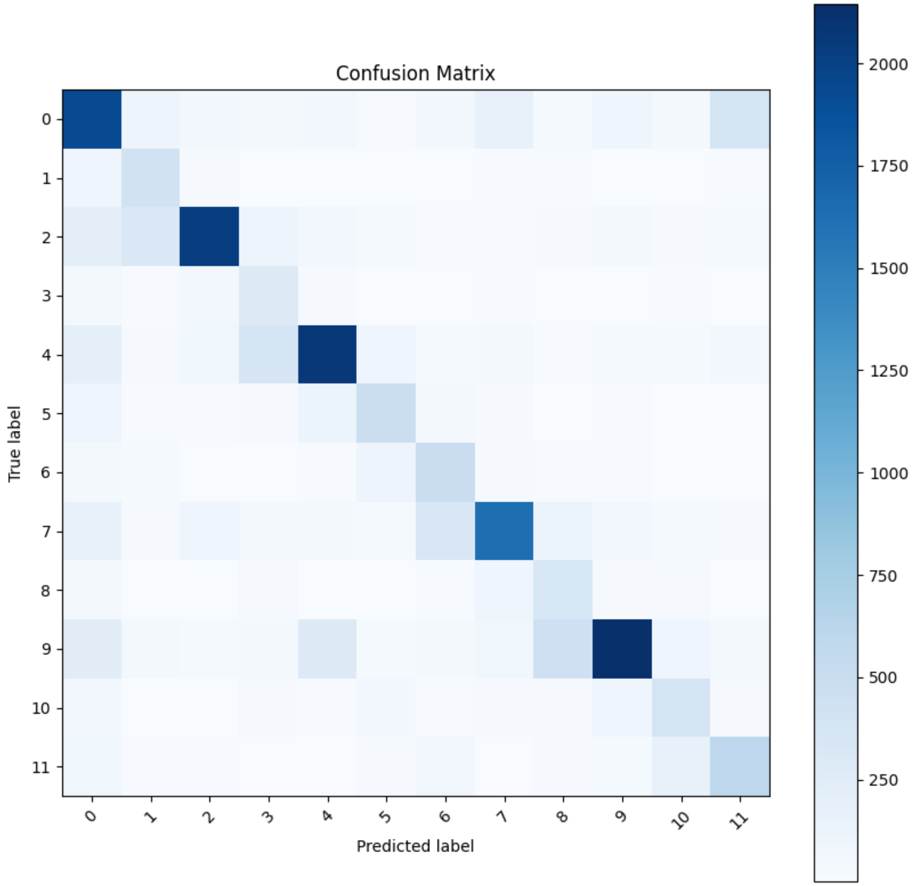
Random chance accuracy:
1/12 →8.3%

Some classifiers for future work:

SVM

Random forest

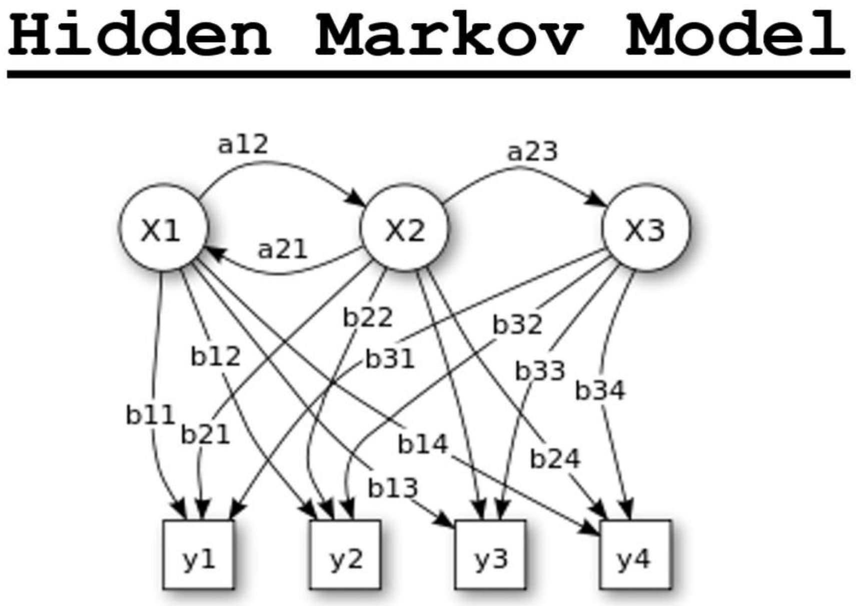
Gaussian Mixture Models



0:C1, **1**:C#1, **2**:D1, **3**:D#1, **4**:E1, **5**:F1, **6**:F#1, **7**:G1, **8**:G#1,**9:** A1, **10**: A#1, **11:** B1

Hidden Markov Model (HMM) for pitch estimation

- Hidden states (X)
- Observations (Y)



Hidden states

Music notes (12 states)

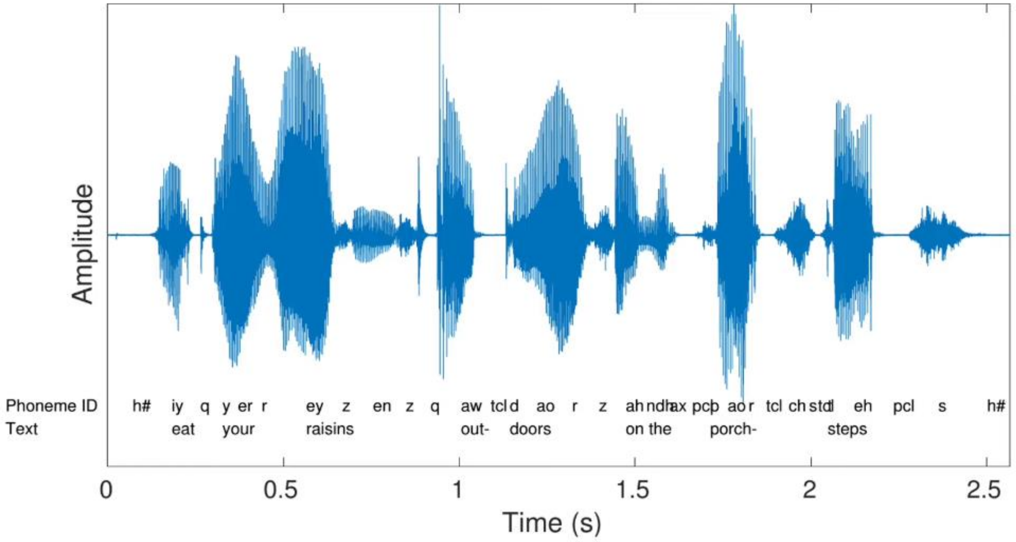
Observations

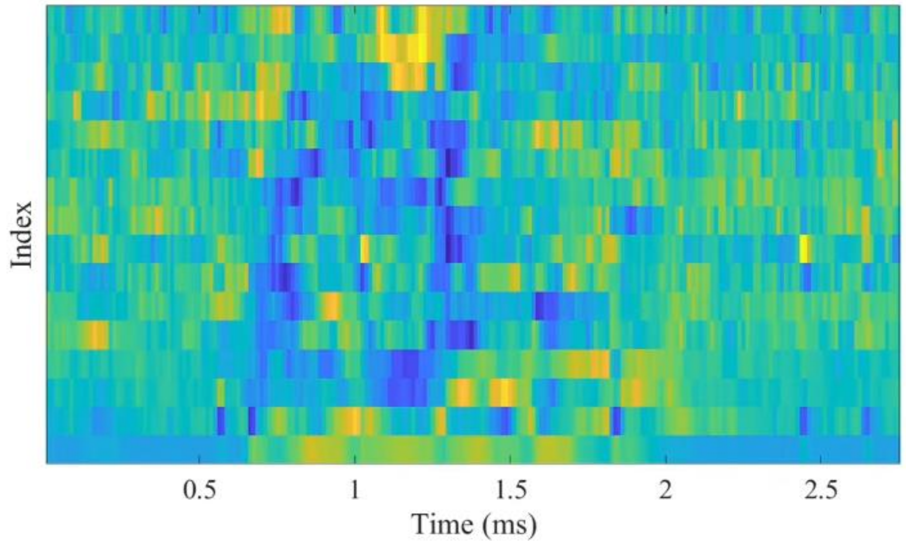
MFCCs (Mel-Frequency Cepstral Coefficients)

$$mfcc_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} \log(x_j) \cos\left(\frac{i\pi}{N}(j - 0.5)\right)$$

How MFCC is computed : summary

1. Windowing
2. Fourier transform (Spectrogram)
3. Triangular filter bank (Mel Spectrogram)
4. Logarithm (Log Mel Spectrogram)
5. DCT (MFCC)





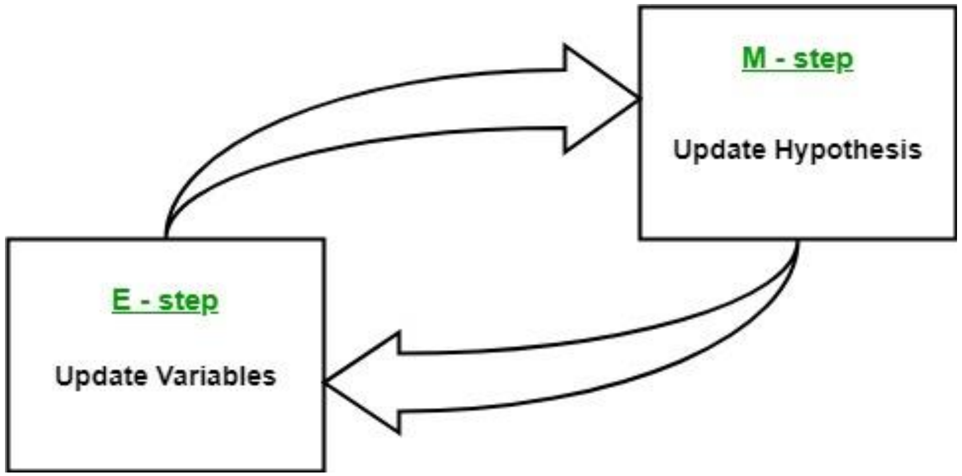
HMM parameters

- Initial probabilities (N°A1/N°Ω)
- Transition probabilities (N°A1->A2/N°A1->Ω)
- Emission probabilities (The Expectation-Maximization (EM) algorithm)

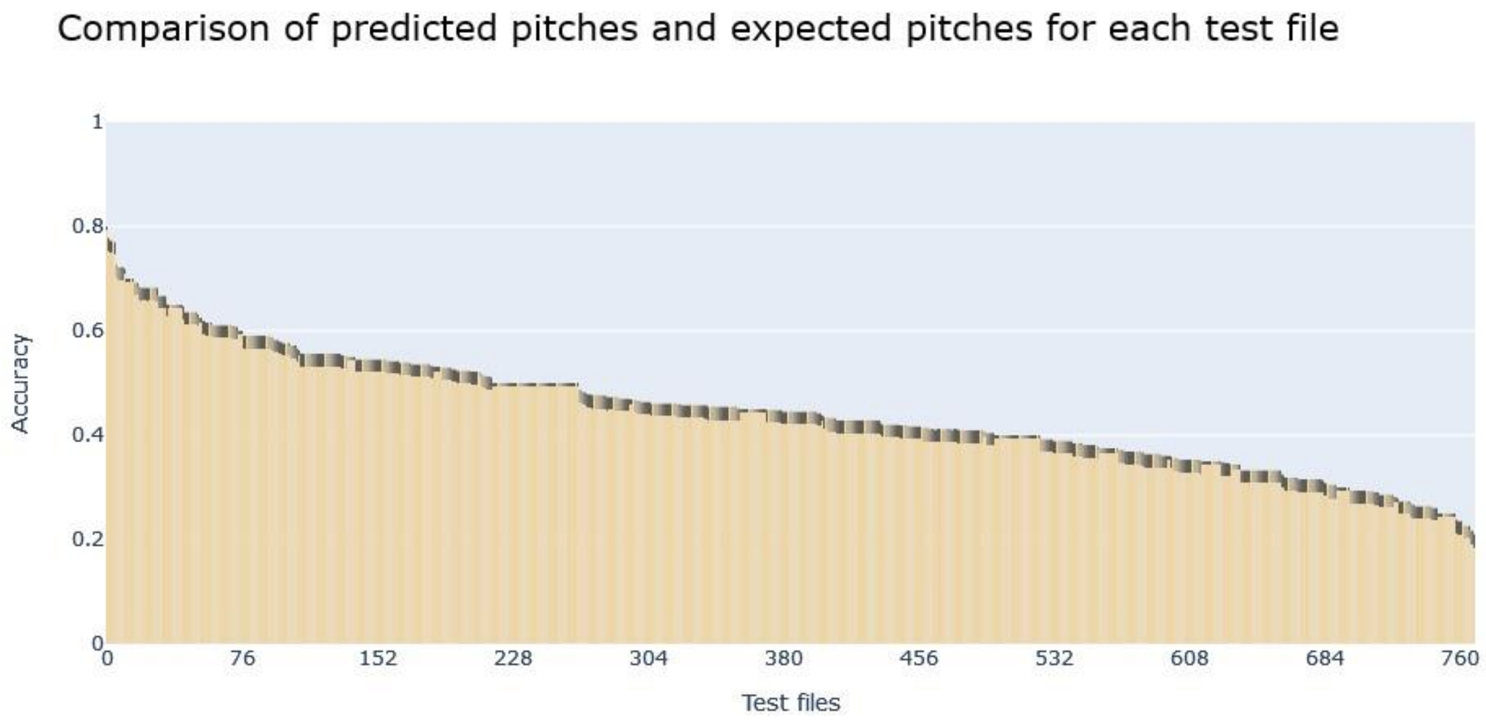
$$\underset{X=X_1,X_2,\dots X_n}{\arg\max} \prod P(Y_i \mid X_i)\ P(X_i \mid X_{i-1})$$

EM Algorithm

- E-step : We estimate a probability of getting a sequence note given a sequence MFCC observations
- M-step : We estimate the emission probabilities that maximizes the likelihood of our MFCCs observation



Accuracy

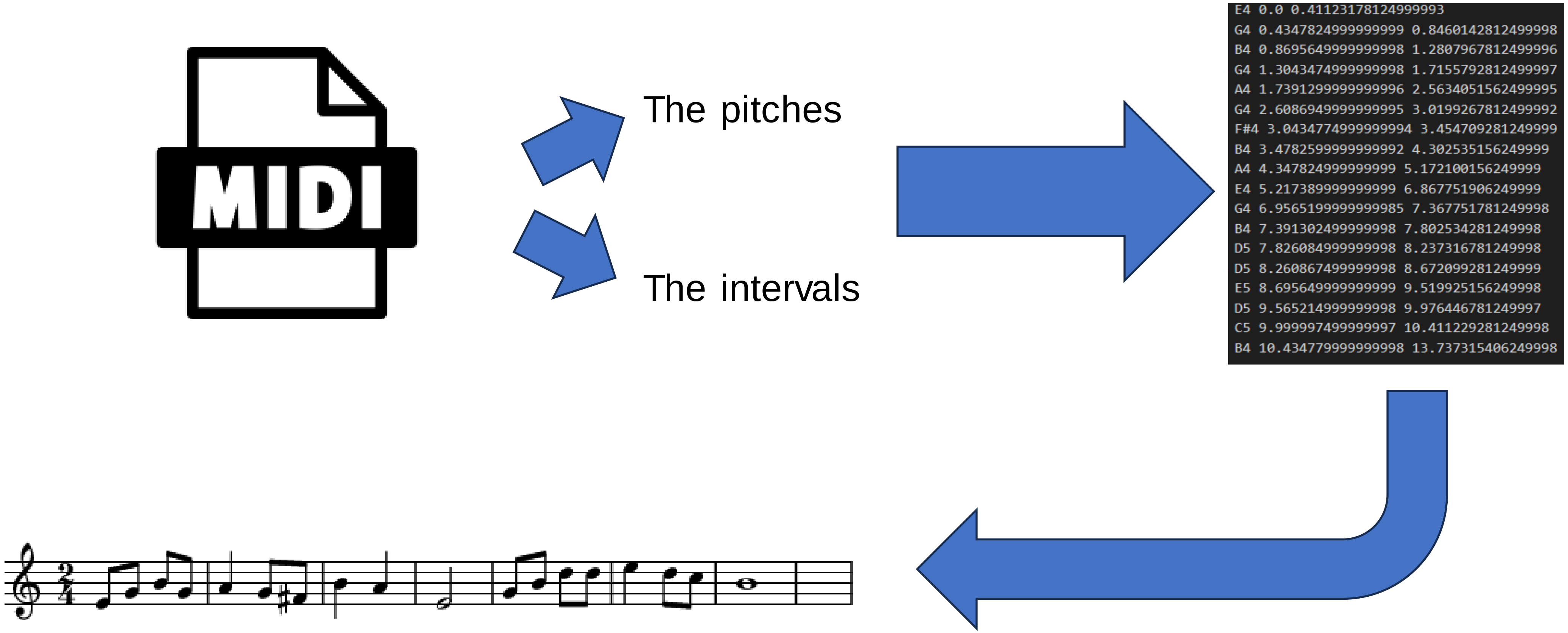


Mean : 0.45
Std : 0.01

- Out of 769 total files, we have 0 file that have been correctly detected, which gives us a success rate of 0.0 for file conversion.
- Out of 19955 total pitches, we have 6886 pitches that have been correctly detected, which gives us a success rate of 0.35 for grade conversion.

Converting MIDI to musical notes

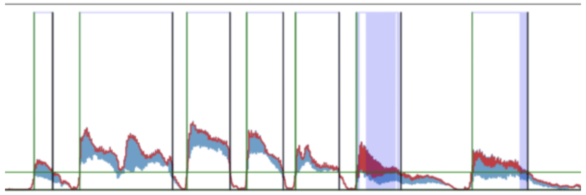
- Music21




```
E4 0.0 0.41123178124999993
G4 0.4347824999999999 0.8460142812499998
B4 0.8695649999999998 1.2807967812499996
G4 1.3043474999999998 1.7155792812499997
A4 1.7391299999999996 2.5634051562499995
G4 2.6086949999999995 3.0199267812499992
F#4 3.0434774999999994 3.454709281249999
B4 3.4782599999999992 4.302535156249999
A4 4.347824999999999 5.172100156249999
E4 5.217389999999999 6.867751906249999
G4 6.9565199999999985 7.367751781249998
B4 7.391302499999998 7.802534281249998
D5 7.826084999999998 8.237316781249998
D5 8.260867499999998 8.672099281249999
E5 8.695649999999999 9.519925156249998
D5 9.565214999999998 9.976446781249997
C5 9.999997499999997 10.411229281249998
B4 10.434779999999998 13.737315406249998
```

CNNs for pitch detection

- Time-Frequency Input Representation:
  - Western Classical music is very geometrical
  - *CQT transform* closely matches music
    - Each semitone can be represented by a fixed number of frequency bins.
    - Each octave occupies same number of frequency bins.
- Our Challenge
  - Imprecise ground truth of onsets and offsets
  - Explored CTC loss based training, DTW, alignment modules, etc
  - Heuristics to correct them – reduced dataset

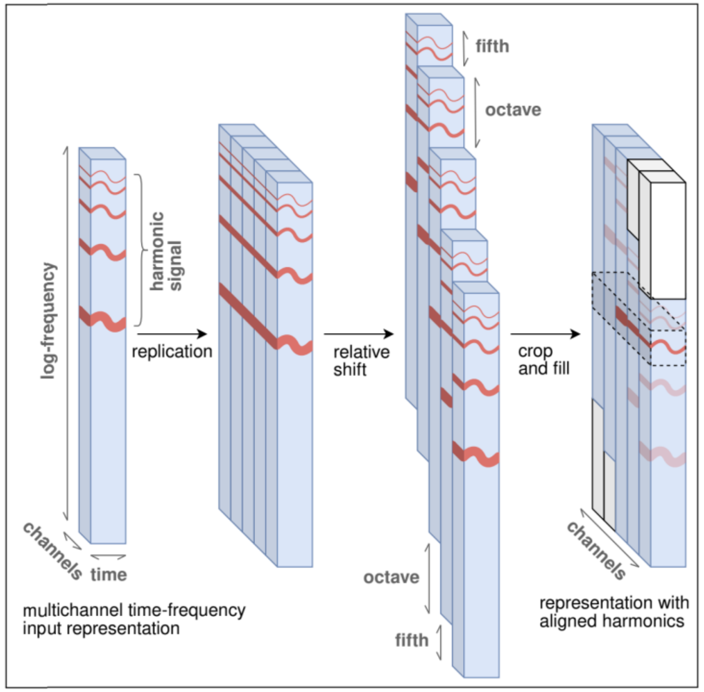


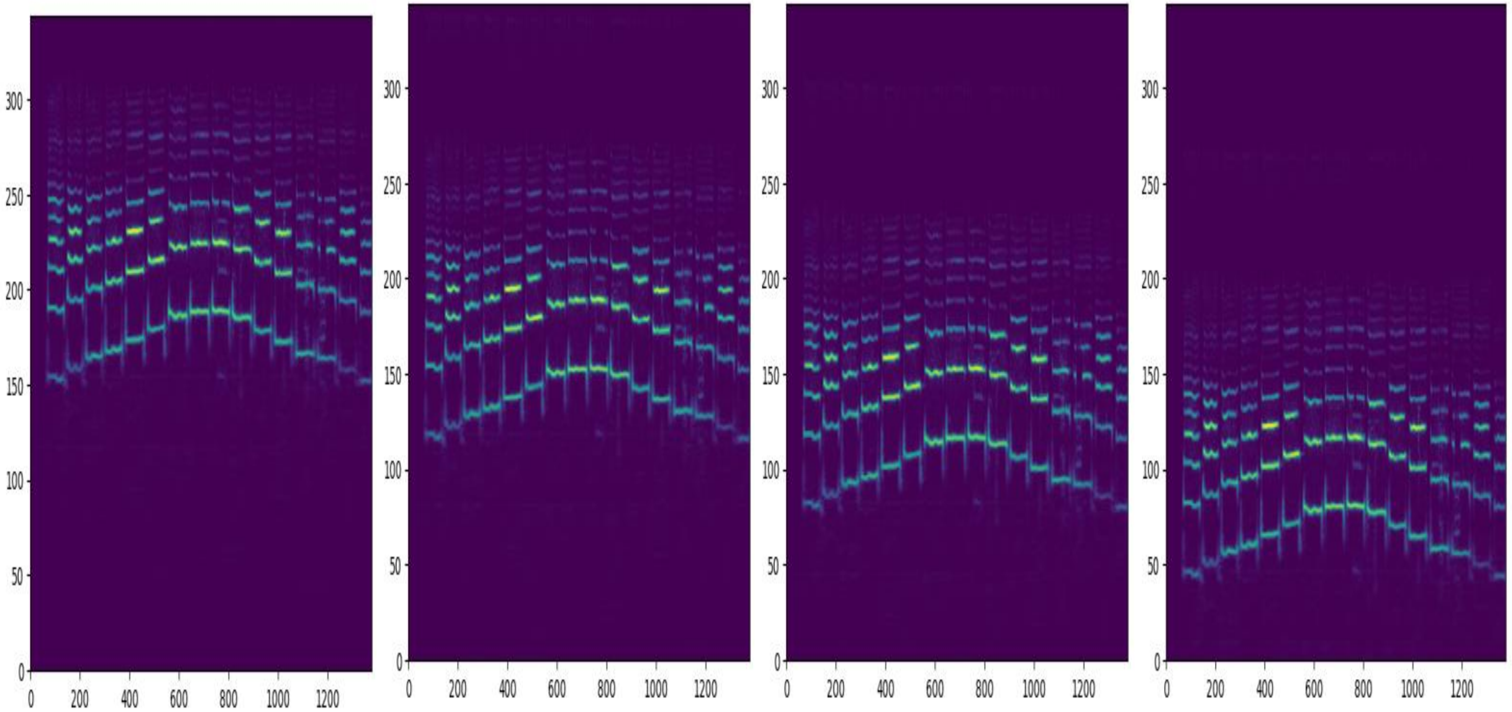
Source: https://www.pinterest.com/pin/345721708865634188/

CNNs for pitch detection – *Harmonic Stacking*

A note hummed by a human has a *fundamental frequency* and its associated *overtones/harmonics*.

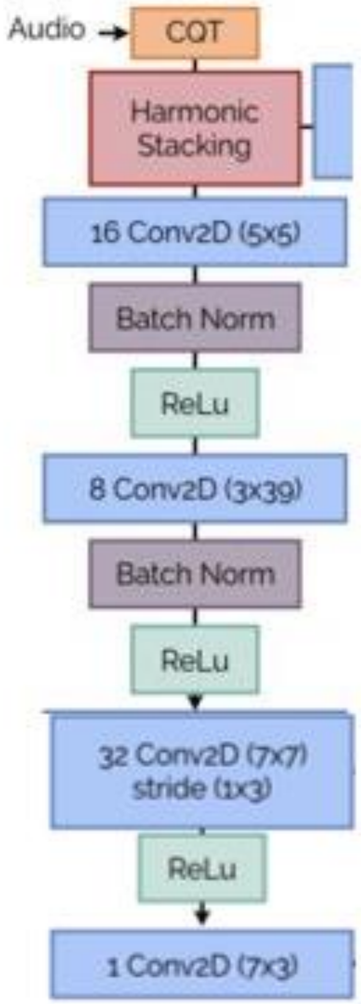
Requires a kernel that can access a large frequency band.





Source: https://www.music-ir.org/mirex/abstracts/2019/BH1.pdf

CNNs for pitch detection – *Model Architecture*

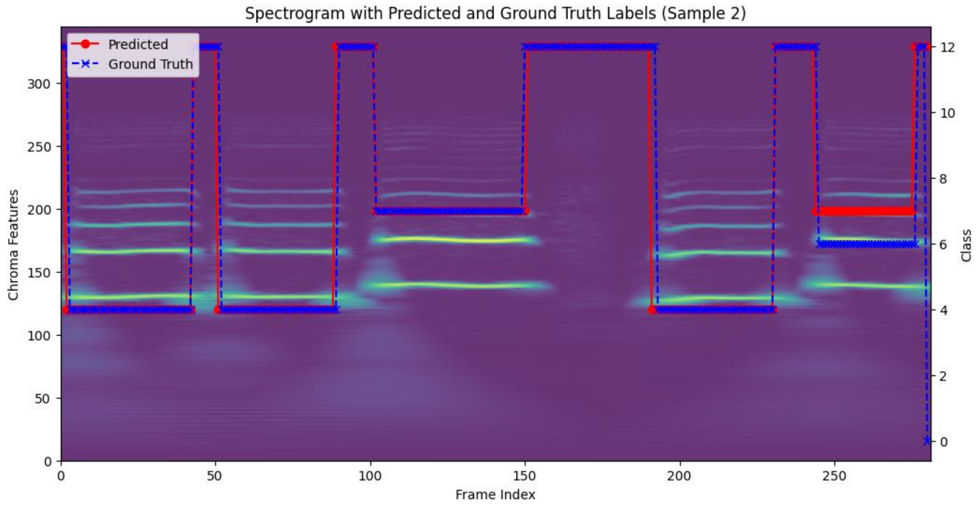


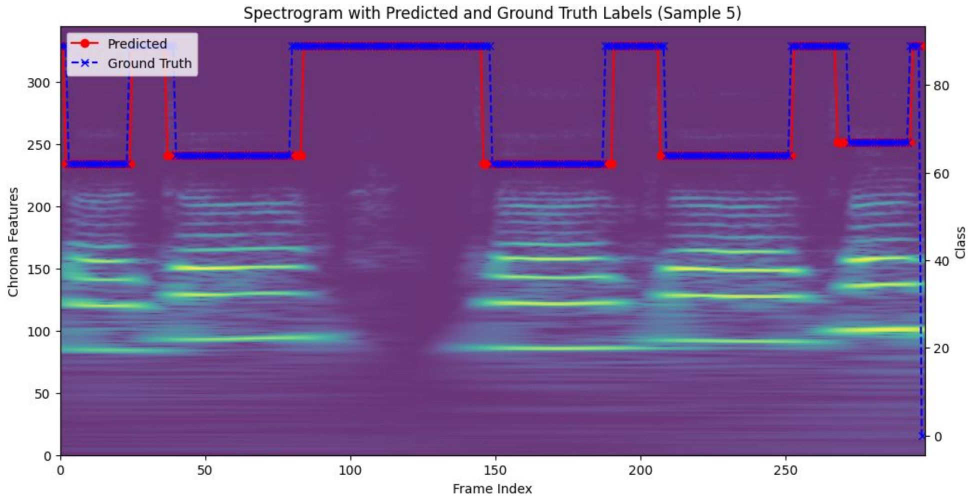
Inspired by Spotify's Basic Pitch model – this architecture employs Harmonic stacking to get access to relevant frequencies in the channel dimension.

Trained with ***Cross Entropy Loss***

Results:

- **Octave Invariant**: Validation accuracy of ~ ***88%***
- **Octave Aware**: Validation accuracy of ~ ***84%***





CNNs for pitch detection – *Future Work*

- Our ground truth is imprecise, employ label smoothing while training the network.

- Ablate different network architecture choices.

- Compare training efficacy of different architectures

- Heuristically cleaning of obtained note onsets, offsets and detected notes.

Conclusion

- We explore a broad range of techniques from classical to deep ML for humming transcription.
- We are working with a novel dataset with no published research based on it so far.
- We contribute by correcting the ground truth onsets and offsets, providing a cleaner dataset for future work

THANK YOU FOR YOUR ATTENTION!

A Hummingbird humming